Mapping UK Biobank to the OMOP CDM: challenges and solutions using the Delphyne ETL framework

Sofia Bazakou* [1], Maxim Moinat[1], Alessia Peviani[1], Anne van Winzum[1], Stefan Payralbe[1], Vaclav Papez[2], Spiros Denaxas[2]

[1] **The Hyve**, Utrecht, The Netherlands * Contact: sofia@thehyve.nl / +31 (0)30 7009713
[2] **University College London**, London, United Kingdom

Background

UK Biobank[1] (UKB) is a large-scale registry containing medical and genetic data from 500,000 consented participants from the UK's general population, aged between 40 and 69 years (Figure 1). UKB is an extraordinary resource for human health research, accessible to approved research initiatives worldwide.

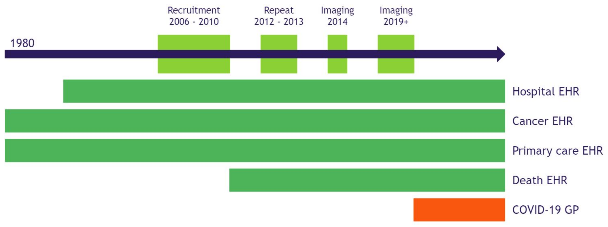

Figure 1. UKB data structure and timeline. The data include baseline assessments (light green), such as surveys, samples, and imaging, electronic health records (EHR) from different sources (dark green), and information on COVID-19 testing (red). Picture adapted from Spiros Denaxas, Professor of Biomedical Informatics, Institute of Health Informatics, University College London.

As part of the European Health Data Evidence Network[2] (EHDEN), The Hyve was contracted by University College London (UCL) to map the UKB data to the OMOP CDM v5.3. The main goal of the collaboration was to make the dataset available for research related to the COVID-19 pandemic. The Hyve implemented the data conversion pipeline, while UCL provided the source data expertise.

The UKB data conversion effort came with several **challenges**:

- ETL development without direct access to the data.
- Mapping of free-text and non-standard ontologies.
- Large heterogeneity of source terms amongst data providers.
- Conversion of a large wide format table to long format.
- Working with an evolving data source.

Methods

The Hyve overcame the lack of direct access to the UKB data by adopting a collaborative Agile-based development approach with UCL (Figure 2). This process made it possible to develop the ETL code relying entirely on **synthetic data**.

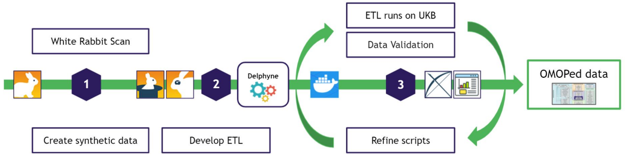

Figure 2. Collaborative Agile development approach using existing OHDSI tools and Delphyne. Step 1: UCL provided a White Rabbit scan report of the UKB data, which The Hyve used to generate synthetic data for ETL development. Step 2: The Hyve performed syntactic and semantic mappings with Rabbit in a Hat and Usagi, respectively, and developed the ETL with Delphyne. Step 3: UCL ran the ETL locally (deployed with Docker) on a UKB data subset, and executed quality checks with Achilles and the Data Quality Dashboard. The Hyve refined the ETL based on feedback before a new iteration. Finally, UCL ran the ETL on the full UKB dataset to produce the complete "OMOPed" data.

Our UKB mapping workflow made use of existing tools from the OHDSI suite and **Delphyne**[3,4], a specialized ETL framework for mapping data to the OMOP CDM, developed internally by The Hyve. Delphyne was particularly helpful in tackling data heterogeneity between UKB healthcare providers, and the mapping of the wide format *baseline* table (500,000x9,000), which both required specific handling logic. Our development approach relied heavily on feedback, which Delphyne helped to provide through detailed logging and summary reports. It also made trivial to extend the CDM model with custom provenance fields, enabling more informative data quality assessments. Finally, Delphyne automated several tasks, such as CDM table creation, vocabulary loading, and mapping of non-standard to standard ontologies, saving time whenever a UKB data or vocabulary update was available. Overall, Delphyne allowed us to build a highly specialized ETL, for maximum mapping coverage and quality.

Conclusion

UKB is an incredible resource for healthcare research. Given its size and complexity, mapping the data to the OMOP CDM model came with several challenges. A powerful and flexible ETL framework such as Delphyne was invaluable in carrying out the conversion effort. Together with existing open-source tools from the OHDSI suite, Delphyne allowed us to perform the conversion without direct access to the UKB data, and to deliver a **high-coverage mapping** (Figure 3). The mapping of UKB data to the OMOP CDM will in turn enable future research, including studies on COVID-19, to build upon our efforts.

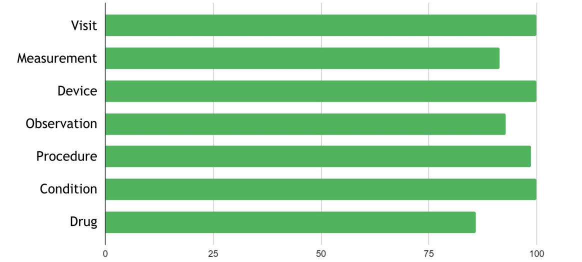

Figure 3. Percentage of UKB source codes mapped to a standard OMOP concept per domain, by record frequency. We achieved a near full mapping coverage for the Visit, Device and Condition domains (>99%), and the lowest mapping coverage for the Drug domain (86%). Note that for the baseline data mapping, we converted a subset of the original variables.

References

[1] https://www.ukbiobank.ac.uk/
[2] https://ehden.eu/
[3] https://www.thehyve.nl/cases/mapping-uk-biobank-to-omop-using-delphyne
[4] https://delphyne.readthedocs.io/en/latest/